Erschienen in: Lobin, Henning/Lemnitzer, Lothar (Hrsg.): Texttechnologie. Perspektiven und Anwendungen. - Tübingen: Stauffenburg, 2004. S. 195-216. (Stauffenburg Handbücher)

Linguistische Korpora

Felix Sasaki und Andreas Witt

1 Einleitung

Die Definition linguistischer Korpora gestaltet sich schwierig. Im Prinzip kann ein Stapel alter Zeitungen oder eine Sammlung handschriftlicher Briefe einer bestimmten Autorin als Korpus angesehen werden. In neuerer Zeit wird allerdings der Begriff Korpus nicht mehr in einer derartig allgemeinen Weise verstanden: Korpora werden als maschinell lesbare, digitalisierte Sprachdaten definiert. Doch auch diese Definition ist noch sehr weit gefasst. Linguistische Korpora im hier behandelten Sinne sind hauptsächlich textuelle Daten, d. h. bereits schriftlich vorliegende Texte oder transkribierte Gespräche. Sie lassen sich abgrenzen von Sammlungen linguistischer, sprachbezogener Daten, bei denen der Text nicht das zentrale Datum ist, wie z. B. Reaktionszeitmessungen in psycholinguistischen Experimenten. Audio- oder Videosignale ohne weitere Informationen fallen ebenfalls nicht unter den Begriff Korpus. Als neuer Korpusbegriff können 'multimodale Korpora' angesehen werden, in denen Verschriftlichungen gesprochener Sprache mit anderen Modalitäten wie Gestik verbunden werden.

Werden linguistische Korpora im Kontext der Texttechnologie betrachtet, ergibt sich eine weitere Verfeinerung des Korpusbegriffs. Zentral für texttechnologische Korpora sind zwei Eigenschaften:

1. Die Texte sind mit Informationen angereichert – Metainformationen oder Informationen, die die verschiedenen linguistischen Beschreibungsebenen (z. B. Morphologie, Syntax, Diskursstruktur) betreffen.
2. Die Informationsanreicherung greift auf texttechnologische Methoden zurück, also Auszeichnungssprachen (vgl. Lobin, in diesem Band) und Annotationskonventionen (vgl. Abschnitt 5, S. 206, und Ule und Hinrichs, in diesem Band).

Der vorliegende Artikel stellt zunächst generelle Fragen (Abschnitt 2): Woher kommen die Korpusdaten, wie wird annotiert, wozu werden die annotierten Korpora verwendet, welche Vorteile bietet die texttechnologische Modellierung von Korpora?

Eine in korpuslinguistischen Arbeiten selten gestellte, aus texttechnologischer Sicht aber zentrale Frage lautet: Wie werden die Korpusdaten repräsentiert? Dieser Frage widmen sich die weiteren Teile dieses Kapitels. Abschnitt 3 (S. 200) stellt die Problematik zunächst anhand der kleinsten textuellen Einheit, dem Schriftzeichen, vor und beschreibt Repräsentationsmethoden, die texttechnologische Standards nutzen: in Bezug auf einzelne Zeichen und mittels Dokumentstrukturierungen. Daraus ergibt sich eine zentrale Rolle für die Standardisierung im Prozess der Informationsanreicherung. Dies betrifft grundlegende Aspekte (Abschnitt 4, S. 205), konkrete Annotationsformate (Abschnitt 5, S. 206) und die Annotation verschiedener Ebenen (Abschnitt 6, S. 209).

2 Generelle Aspekte von Sprachkorpora

2.1 Woher kommen die Daten?

Eine häufig genutzte Quelle zur Korpusbildung sind Zeitungen. Für das Deutsche kann hier exemplarisch die am Institut für deutsche Sprache (IDS) in Mannheim erstellte Sammlung der Cosmas-Korpora (*Corpus Storage, Maintenance and Access System*) stehen. Sie besitzt derzeit einen Umfang von 1 736 Millionen Wörtern und wurde hauptsächlich aus Zeitungstexten gespeist. Eine andere Quelle sind literarische Werke. Insbesondere für multilinguale Übersetzungskorpora eignen sie sich gut, wenn von einem Werk Übersetzungen in vielen Sprachen vorliegen. Im Rahmen des Projektes Multext-East (vgl. Erjavec et al., 1998) wurden beispielsweise Übersetzungen von George Orwells *1984* parallel annotiert und bis zur Satzebene automatisch aligniert. Von Relevanz für die Rechtswissenschaften sind parallele Korpora von Texten des „Official Journal of the European Community", wie sie im Multext-Projekt anhand der CES-Konventionen[1] annotiert wurden.

Einige Korpora erheben den Anspruch, eine Sprache in ihrem Gebrauch zum Zeitpunkt der Korpuserstellung zu repräsentieren. Ein bekanntes Beispiel ist das British National Corpus (BNC), welches mehr als 100 Millionen Wörter geschriebener und gesprochener Sprache enthält (vgl. Aston und

[1] Zu den CES-Konventionen vgl. Ule und Hinrichs (in diesem Band).

Burnard, 1998). Um die Repräsentativität der Daten gewährleisten zu können, sind möglichst viele Varietäten des Englischen enthalten. Auch für das Deutsche wird an derartigen Korpora gearbeitet, etwa am deutschen Referenzkorpus (DeReKo) oder im Rahmen des Projektes „Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts" (DWDS) am „Referenzcorpus für die deutsche Sprache des 20. Jahrhunderts"[2].

Die letzte hier vorgestellte Datenquelle sind Korpora, die bewusst einen Sprach- oder Phänomenbereich fokussieren. Das *Child Language Data Exchange System* (CHILDES, vgl. McWhinney, 2000) etwa enthält ausschließlich Korpora, die den kindlichen Spracherwerb dokumentieren. Die Daten können anhand bestimmter Konventionen annotiert und mit eigens dafür entwickelten Tools analysiert werden. Aus dem Bereich der automatischen Übersetzung kann das Projekt Verbmobil genannt werden, in dessen Rahmen deutsche, englische und japanische Korpora erstellt wurden (vgl. Burger et al., 2000). Diese Korpora entstanden unter bestimmten Szenariovorgaben (Terminabsprachen, Reiseplanung), um so die Qualität der Übersetzung zu erhöhen.

2.2 Wozu wird annotiert?

Korpora sind für manche linguistische Disziplinen unverzichtbar. Soziolinguistische Studien etwa, die Korrelationen zwischen sprachlichen Eigenschaften und Parametern wie Alter, Geschlecht oder sozialer Herkunft untersuchen, arbeiten häufig mit Korpora. Auch andere linguistische Gebiete wie Grammatik, Semantik, Pragmatik oder Stilistik nutzen teilweise Korpora.[3]

Die Arbeit mit und der Nutzen von Korpora wurde nicht von allen linguistischen Schulen akzeptiert. Insbesondere die frühe generative Grammatik betrachtete Korpora als Sammlung 'zufälliger Erscheinungen', da sie nur Instanzen des Sprachgebrauchs beinhalten. Für die linguistische Theoriebildung waren diese 'zufälligen Erscheinungen' unbrauchbar, statt dessen wurden Äußerungen von kompetenten Sprechern gebildet und hinsichtlich ihrer Grammatikalität bewertet.

In neuerer Zeit hat sich dieses Bild gewandelt. Korpora finden in den verschiedensten grammatischen Schulen Verwendung, um die Existenz bestimmter syntaktischer Konstruktionen zu überprüfen oder unbekannte Konstruktionen auffinden zu können.

Vgl. `http://www.ids-mannheim.de/dereko/` und `http://www.dwds.de/` für weitere Informationen.

Vgl. McEnery und Wilson (1996) für einen ausführlichen Überblick.

In computerlinguistischen Anwendungen können mit Korpora zum einen Ressourcen wie etwa Lexika entwickelt werden (vgl. hierzu Lemnitzer und Wagner, in diesem Band). Die Korpora bilden die empirische Grundlage für syntaktische Merkmale lexikalischer Einträge und mögliche Subkategorisierungen. Zum anderen bieten sie Trainings- und Evaluationsmaterial für sprachtechnologische Anwendungen. Zu nennen sind hier Korrekturprogramme, Dialogsysteme, Text Mining, Systeme zur maschinellen Übersetzung etc. Am Beispiel morphologischer Tagger sei eine derartige Einsatzweise für Korpora vorgestellt: Ein verlässlich annotiertes Korpus wird zunächst mit verschiedenen Methoden (siehe Abschnitt 2.3) analysiert; die daraus resultierenden Informationen bilden die Grundlage zur automatischen Annotation eines unbearbeiteten Korpus. Die Qualität des automatischen Annotationsverfahrens kann anhand eines unbearbeiteten Korpus getestet werden, die Ergebnisse dieser Evaluation fließen wiederum in seine Verbesserung mit ein.

2.3 Wer annotiert was und wie?

Im Prozess der Annotation wird der Text mit Informationen angereichert. Diese Informationen können als Metadaten (vgl. Schmidt, in diesem Band) die Gesamtheit des Textes betreffen oder auch einzelne Einheiten. Bei linguistischen Korpora sind es die Einheiten linguistischer Analyse, die Gegenstand der Annotation sein können: orthographische Informationen (vgl. Abschnitt 3), Wortarten (engl. *part of speech*, POS), Lemmatisierungen, syntaktische Strukturen, semantische Informationen, diskurs- und textlinguistische Einheiten (z. B. anaphorische Beziehungen), phonetische und prosodische Transkriptionen.

Grundsätzlich können die Sprachdaten auf zwei unterschiedlichen Wegen mit linguistischen Informationen ausgezeichnet werden, zum einen automatisch, zum anderen manuell. Da beide Methoden Vor- und Nachteile besitzen, finden sich im praktischen Einsatz auch Kombinationen beider Verfahren.

Die automatische Annotation kann regelbasiert verlaufen wie in der *Constraint Grammar* (Karlsson et al., 1995); stochastische Modelle wie z. B. HMM (*hidden Markov models*, vgl. Mehler, in diesem Band) bieten einen anderen Weg. Ist die Datengrundlage ein beliebiger Text, kann die Häufigkeit komplexer Strukturen die automatische Annotation selbst einfacher Einheiten erschweren. Für diese Fälle wird oft ein partielles, regelbasiertes oder stochastisches Parsing eingesetzt, dass sich auf einfache, d. h. nicht verschachtelte Nominalphrasen beschränkt. Je nachdem, welche Einheiten zu

annotieren sind, ist die automatische Annotation mehr oder weniger erfolgreich. Es ist bisher nicht möglich, eine sowohl fehlerfreie als auch vollständige linguistische Annotation maschinell zu erreichen.

Der offensichtlichste Nachteil der manuellen Auszeichnung ist der hohe Zeitaufwand für ihre Erstellung, die Qualität ist jedoch höher. Aufgrund der Grenzen automatischer Verfahren sind jedoch manuelle Annotationen zumindest teilweise unumgänglich. Allerdings ist das Ziel der Fehlerfreiheit nicht unbedingt erreichbar. So wählen z. B. verschiedene Personen eventuell verschiedene Antezedenten von anaphorischen Konstruktionen. Nichtsdestotrotz gilt, dass die Qualität der von Menschen durchgeführten Annotation weitaus höher ist als die maschinelle. Entsprechend geben Vergleiche von automatisch und manuell annotierten Texten Aufschluss über die Leistungsfähigkeit der maschinellen Verfahren. Die manuellen Annotationen können getestet werden, indem verschiedene Annotatoren die gleichen Daten bearbeiten und die Ergebnisse verglichen werden.

Bei der Annotation kommen verschiedene Tools zum Einsatz, wie z. B. die im MATE-Projekt entwickelte Workbench (McKelvie et al., 2001). Einige Tools werden speziell für bestimmte Phänomene oder linguistische Theorien entwickelt. Manche Tools konzentrieren sich auf eine bestimmte Art von Korpora, wie z. B. multimodale Korpora (vgl. Milde und Gut, 2002).

2.4 Linguistische Korpora und Texttechnologie

Die Anwendung texttechnologischer Methoden bietet für linguistische Korpora verschiedene Vorteile. Zunächst ist die Wiederverwertbarkeit von Ressourcen zu nennen. Nutzt ein Korpus verbreitete Auszeichnungssprachen und Annotationskonventionen, so kann es mit vertretbarem Aufwand verschiedenen Zwecken oder computerlinguistischen Systemen zugeführt werden. Zudem bieten weit verbreitete Auszeichnungsformate wie HTML die Aussicht, große Datenmengen schnell zu erschließen (vgl. Abschnitt 5.1, S. 206). Auch die Korpuserstellung wird durch texttechnologische Methoden erleichtert. Die Daten verschiedener Tools, welche auf bestimmte sprachliche Ebenen – z. B. Annotation von Morphologie *vs.* Prosodie – spezialisiert sind, können durch Auszeichnungssprachen als gemeinsame Basis miteinander verknüpft werden. Bestimmte Korpustypen wie solche multilingualen Korpora, in denen unterschiedliche Schriftsysteme parallel enthalten sind, werden erst möglich durch Standardisierungen auf der grundlegenden, d. h. der Zeichenkodierungsebene (siehe Abschnitt 3, S. 200).

Bedeutsam nicht nur für texttechnologische Korpora ist die Repräsentation von Informationen unterschiedlicher Beschreibungsebenen – einzel-

ne Zeichen, Morpheme, Wörter, Phrasen etc. Texttechnologische Methoden bieten die Möglichkeit, die Ebenen in der nötigen Granularität zu kodieren und miteinander in Beziehung zu setzen, je nach gewähltem Annotationsformat in mehr oder weniger variabler Weise (vgl. Abschnitte 5 und 6).

Ein selten problematisiertes Thema ist die grundlegende Ebene der Korpusbildung, das einzelne Zeichen. Dies mag daran liegen, dass die Zeichenrepräsentation bei einsprachigen Korpora mit europäischen Sprachdaten wenig Hindernisse bereitet. Sobald jedoch multilinguale Daten oder z. B. asiatische Schriftsysteme im Korpus enthalten sein sollen, sind bestimmte Verfahren der Zeichenrepräsentation unabdingbar. Sie sind Thema des folgenden Abschnittes.

3 Multilingualität, Sprache, Schrift, Codierung

3.1 Basiseinheiten: Zeichen und Glyphen

Ein Zeichen, engl. *character*, lässt sich auffassen als „The smallest component of written language that has a semantic value [...]" (ISO:10646, 2000). Je nach Sprache können hiermit sehr verschiedene Einheiten gemeint sein, etwa phonembezogene Zeichen wie in den lateinbasierten Schriften, Silben wie in den japanischen Hiragana oder piktographische Zeichen wie bei den chinesischen Kanji.

Ein Glyph ist „a recognizable abstract graphic symbol which is independent of a specific design" (ISO:9541-1, 1991). Das Verhältnis von Glyphen zu Zeichen ist nicht immer eindeutig. Ein Zeichen kann durch einen oder mehrere Glyphen repräsentiert werden, wie das Zeichen 'Ä' singulär oder durch die beiden Glyphen 'A + Umlautmarkierung'. Ein einzelnes Glyph kann eine Sequenz von Zeichen darstellen, wie bei der Ligatur 'æ'. Ein Zeichen kann je nach Kontext mit verschiedenen Glyphen dargestellt werden, wie im Falle der arabischen Schrift, die Glyphen in Abhängigkeit von ihrer Position im Wort differenziert. Schließlich kann ein Glyph verschiedene Zeichen repräsentieren wie das Glyph 'A' ein Zeichen im lateinischen, griechischen oder kyrillischen Alphabet. Glyphen können mit bestimmten Visualisierungen, *glyph images* verknüpft werden. Diese werden zusammengefasst in einem Font, z. B. Arial 'A' oder Courier New 'A'.

Bei multilingualen, linguistischen Korpora bestimmen der Zweck der Korpuserstellung, die Analysemethodik oder die Präsentationsvorgaben die

Wahl des Schriftsystems, also des Zeichenvorrates[4]. Ein Korpus zur Analyse gesprochener Sprache kann zunächst sprachunabhängig bleiben und auf das IPA-Inventar (International Phonetic Alphabet) zurückgreifen. Ein textbezogener Bearbeitungsprozess wie Tagging benötigt jedoch Daten, die Glyphen hinreichend genau differenzieren, je nachdem welche Einheit analysiert werden soll. Sprachtypologische Untersuchungen oder die Verschriftlichung 'schriftloser' Sprachen können zudem nicht auf etablierte Symbolinventare zurückgreifen. Oft wird eine lateinbasierte Verschriftlichung verwendet. Um sprachspezifische Eigenschaften wie lexikalisierte Tonverläufe darstellen zu können, reicht weder das IPA-Inventar noch die lateinbasierte Form aus, so dass neue Zeicheninventare entwickelt werden oder das lateinische ergänzt wird (Musgrave, 2001). Bestimmte Analysen schließlich sind auf Grund der Granularität eines Schriftsystems nicht adäquat durchführbar, etwa wenn das vorherrschende Schriftsystem einer Sprache syllabisch strukturiert ist und für eine morphologische Strukturierung nicht fein genug differenziert. Eine Lösung wäre die Verwendung einer lateinbasierten Verschriftlichung, mit dem Nachteil, auf die Präsentation des Korpus in dem gebräuchlichen Schriftsystem verzichten zu müssen.

Im Folgenden werden Verfahren vorgestellt, mit denen sich verschiedene Schriftsysteme in einem Korpus parallel kodieren und zueinander in Beziehung setzen lassen. Gegenstand ist zunächst die Kodierung von Zeichen als singuläre, numerische Einheiten und anschließend die Sprach- und Schriftspezifizierung auf der Ebene von Dokumentauszeichnungen bzw. -strukturierung.

3.2 Kodierung mittels singulärer Einheiten

Zur Definierung einer textuellen, maschinenlesbaren Kodierung müssen folgende Schritte unternommen werden (vgl. Whistler und Davis, 2000):

1. Die Menge der zu kodierenden Zeichen wird ausgewählt.
2. Jedes Zeichen wird mit einem numerischen Identifikator, engl. *code point* assoziiert. Das Ergebnis ist ein kodierter Zeichensatz, engl. *coded character set* (CSS). Der Großbuchstabe 'A' erhält etwa den numerischen Wert '65'.
3. Die grundlegende Dateneinheit wird bestimmt, etwa ein Byte oder ein Doppelbyte, und das CSS auf diese Einheit abgebildet. Das 'A' etwa kann durch ein Byte mit dem entsprechenden Wert '65' repräsentiert werden.

[4] Vgl. Gippert (1999b), insbesondere Gippert (1999a) für eine ausführliche Behandlung des Themas.

4. Ein Serialisierungschema oder *character encoding scheme* (CES) wird entwickelt, in welchem die Bytereihenfolge definiert wird. Dies ist relevant, wenn Zeichenvorräte kodiert werden müssen, die nicht allein durch ein Byte darstellbar sind.

Ein Zeichensatz, engl. *character set*, verknüpft schließlich CSS und CES. Die Namen gebräuchlicher Zeichensätze werden definiert in Übereinstimmung mit der *Internet Assigned Numbers Authority* (IANA, vgl. `http://www.iana.org/assignments/character-sets`). Im Folgenden wird die Problematik der Zeichensätze in linguistischen Korpora hinsichtlich der Auswahl der Zeichen und ihrer maschinenlesbaren Repräsentation behandelt.

Auswahl der Zeichen

Viele Zeichensätze deklarieren Zeicheninventare für bestimmte Sprach- und Kulturräume. Um in unterschiedlich lokalisierten Softwareumgebungen eine möglichst breite Verwendbarkeit der gleichen textuellen Daten zu gewährleisten, überlappen sich große Bereiche: Der Großbuchstabe 'A' etwa ist bei ASCII oder den Zeichensätzen der ISO-8859-Familie (Latin 1, 2, ...) immer identisch numerisch bezeichnet. Zusätzlich zum ASCII-Vorrat definiert ISO:8859-1 (1998) Zeichen für westeuropäische Sprachen (Deutsch, Englisch, Französisch etc.) oder ISO:8859-2 (1999) Zeichen für mitteleuropäische und slawische Sprachen (Kroatisch, Polnisch, Rumänisch etc.).

Bei multilingualen Korpora stößt diese Vorgehensweise schnell an Grenzen: Die Verarbeitung der Texte ist schwer zu realisieren, wenn nicht alle notwendigen Zeichen gleichzeitig zur Verfügung stehen. Der auch in XML angewandte ISO-Standard ISO:10646 (2000), definiert durch das Unicode-Konsortium, scheint hier Abhilfe zu schaffen. Unicode, hier auf den Standard in der Version 3.0 bezogen, weicht in zweierlei Hinsicht von den angesprochenen Zeichensätzen ab. Zum einen sind Überlappungen in der Zeichendefinition nicht vorgesehen. Zum anderen sind die Zeichen nicht nach Sprach- oder Kulturräumen geordnet, sondern nach „Skripten". Die Skripte umfassen Zeicheninventare weitgehend unabhängig davon, welche regionale und sprachliche Verwendung sie finden. Im so genannten „Basic Multilingual Plane" (BMP), das bis zu 65536 Zeichen umfasst, werden weitverbreitete Skripte definiert.

Die Schwierigkeit des Unicode-Ansatzes besteht darin zu entscheiden, ob ein Zeichen einen eigenen numerischen Identifikator erhalten soll oder ob es als Glyph einzustufen ist. Die Differenzierung in Sprachen oder Sprachräume, historische oder regionale Varianten wird auf der Ebene der numerischen

Identifikatoren nicht berücksichtigt. Probleme erwachsen auch bei der arabischen Schrift, deren kontextabhängige Schreibung von Unicode aus dem Bereich der Repräsentation zur Zeichendarstellung verschoben wird.

Es gibt verschiedene Lösungsansätze dieses Problems. Unicode stellt einen Bereich zur Verfügung, in dem der Benutzer selbst Zeichen definieren kann. Die Verwendung dieses Bereiches hat den Nachteil, dass die softwareübergreifende Verarbeitung nicht mehr gewährleistet ist. Eine andere Möglichkeit bietet der Übergang auf die Ebene der Dokumentauszeichnung und -strukturierung (vgl. Abschnitt 3.3, S. 203).

Maschinenlesbare Kodierung der Zeichen

In Unicode können Zeichen als UTF-8 oder UTF-16 kodiert werden. UTF-8 verwendet Sequenzen von einem bis zu drei Byte und hat den Vorteil, bei Dokumenten mit entsprechend beschränkten Zeichen kompatibel zu ASCII zu sein; eine Konvertierung entfällt und der Speicherbedarf bleibt gering. Sind in einem Dokument jedoch viele außerhalb des ASCII-Bereiches liegende Zeichen, insbesondere nicht lateinbasierte Zeichen, enthalten, steigt der Speicherbedarf erheblich. Ein UTF-16 Dokument hingegen verwendet Doppelbytesequenzen, unabhängig davon, ob es nur ASCII Zeichen enthält oder nicht. Die Beantwortung der Frage, ob UTF-8 oder UTF-16 geeignet ist, hängt also davon ab, ob ausschließlich lateinbasierte Zeichen kodiert werden müssen oder ob in hohem Maße darüber hinaus gegangen wird.

Die Wahl von Zeichensatz und maschinellem Kodierungssystem hat große Auswirkungen auf die Korpusbearbeitung. Je nachdem, ob Zeichen einen eigenen numerischen Identifikator besitzen oder nur als graphische Variante anderer Zeichen deklariert sind, sind sie einfach durch Such- oder Sortieralgorithmen erfassbar oder nicht. Zudem sind in Unicode sprachspezifische Sortierreihenfolgen auf Grund der sprachunabhängigen Konzeption schwer zu realisieren.

3.3 Kodierung mittels Dokumentauszeichnung und -strukturierung

Dokumentauszeichnungen können genutzt werden, um in der aktuellen Kodierung nicht enthaltene Zeichen zu repräsentieren oder verschiedene Zeichen miteinander in Beziehung zu setzen. Die TEI beschreibt eine „writing system declaration" (WSD), d. h. ein logisch vom zu annotierenden Text unabhängiges Dokument, in dem zusätzliche Informationen über ein Zeichen

repräsentiert werden können (Sperberg-McQueen und Burnard, 1994, Kapitel 25). Im folgenden Beispiel wird die Transliteration eines griechischen 'A' in Lateinumschrift mit dem Zeichen aus dem nativen griechischen Alphabet, repräsentiert durch den Wert '03B1', verknüpft:

```
<character class='lexical'>
  <form string='A' entityStd="agr" ucs-4='03B1'>
    <desc>Greek small letter alpha</desc>
  </form>
</character>
```

Mit diesem Verfahren ergibt sich die Möglichkeit, verschiedene Glyphen zu einem abstrakten Zeichen zu relationieren.

Zeichenkodierungsinformationen nicht in ein zusätzliches Dokument wie die WSD, sondern direkt in die Annotation zu integrieren, ist ein Weg, der z. B. in dem XHTML-Modul *ruby* verwendet wird (vgl. Suignard et al., 2001).

Das folgende Beispiel umfasst einen Text in japanischer Originalschrift und eine Transliteration:

```
<ruby>
  <rbc><rb>?</rb><rb>?</rb><rb>?</rb></rbc>
  <rtc><rt>ka</rt><rt>e</rt><rt>ru</rt></rtc>
</ruby>
```

Das Element `ruby` enthält zunächst ein Element `rbc`, „ruby base container", in welchem der Basistext in einzelne Segmente `rb` zergliedert ist. In dem Element `rtc` („ruby text container") wiederum sind die Transliterationen der Segmente in phonetischer Umschrift enthalten, die zum Basistext aligniert werden. Die Ausgabe in einem Browser ist in Abbildung 3.3 wiedergegeben.

XML bietet über das Attribut `xml:lang` die Möglichkeit, für bestimmte Bereiche von Dokumenten Sprachen anzugeben. So können z. B. chinesische und japanische Glyphen durch die Attributwerte `zh` und `jp` nach ISO:639 (1998) differenziert werden. Die Dokumentauszeichnung kann auch dazu dienen, die textuellen Daten feiner zu segmentieren als es das gebräuchliche Schriftsystem zulässt. Für die angesprochene morphologische Strukturierung syllabischer Schriftsysteme müssen in der Dokumentauszeichnung Informationen enthalten sein, mit denen sich ein geeignetes Schriftsystem – z. B. IPA oder eine Lateinumschrift – generieren lässt und die syllabischen Daten morphologisch neu segmentiert werden können. Eine derartige Kodierung mittels Dokumentauszeichnungen birgt allerdings die Gefahr,

Abbildung 1: Darstellung der Ruby-Annotation in einem Browser

anwendungs- oder domänenspezifisches und somit nicht wiederverwendbares Material zu erzeugen. Um dem Prinzip 'für jede Annotation eine Transformationskomponente' zu entgehen, ist die Standardisierung auch auf der Strukturierungsebene von Dokumenten ein wichtiger Aspekt. Sie ist Thema der folgenden Abschnitte.

4 Informationsanreicherung

Wie bereits gesehen, bestehen linguistische Korpora aus den textuellen Daten und Informationen über die Daten. Zum einen betreffen diese zusätzlichen Informationen das gesamte Dokument, z. B. den Brief, das Buch oder das transkribierte Gespräch, zum anderen annotieren sie Segmente der Daten bezüglich der Zugehörigkeit zu bestimmten Umgebungen (z. B. „der Textteil gehört zu einer Äußerung, einem Satz oder einem Buch") bzw. markieren singulär auftretende Ereignisse (z. B. Pausen in gesprochener Sprache oder Zeilenumbrüche und Seitenwechsel).

Der erste Aspekt ist bereits von Schmidt (in diesem Band) thematisiert worden, so dass hier der zweite Punkt, die direkte Annotation der Daten, fokussiert werden soll.

Eine zentrale, am Beginn des Korpusaufbaus stehende Entscheidung betrifft die Wahl des Repräsentationsformates. Es besteht weitgehend Konsens darüber, dass hierfür nach Möglichkeit standardisierte Annotationskonventionen zu verwenden sind. Allerdings ist nicht unumstritten, was genau unter Standardisierung zu verstehen ist. Werden die in existierenden Korpora Verwendung findenden Auszeichnungen aus einer texttechnologischen Perspektive betrachtet, so lassen sie sich in zwei Gruppen einteilen: Sie verwenden die in Lobin (in diesem Band) dargestellten standardisierten Auszeich-

nungssprachen oder sie verwenden sie nicht. Eine derartige Klassifikation soll jedoch keineswegs den Schluss nahe legen, dass alle XML-basierten Annotationen als standardisiert einzustufen sind. Zweifelsohne stellt XML eine Standardsyntax zur Verfügung, jedoch können innerhalb dieser Syntax sehr unterschiedliche Dokumenttypen definiert werden. Nur wenige der in diesem Rahmen definierten, potentiell unbegrenzt großen Anzahl von Annotationskonventionen erreichen den Status eines Standards für die Auszeichnung linguistischer Korpora. Auf der anderen Seite gibt es Korpora, z. B. das bereits angesprochene CHILDES-Korpus, deren Annotationskonventionen nicht auf XML basieren und die dennoch einen weithin akzeptierten Standard bilden. Allerdings besitzen derartige Standards eine längere Entwicklung, deren Beginn in die Vor-XML-Zeit fällt. Neuere korpuslinguistische Annotationsformate basieren auf XML.

Der Wunsch, standardisierte Auszeichnungsformate zu verwenden, steht oft in Opposition zu einer anderen Anforderung. Korpora werden meist zur Untersuchung bestimmter Forschungsfragestellungen oder im Kontext einer speziellen Theorie aufgebaut. Hieraus erwächst der Bedarf, z. T. sehr spezielle Phänomene auszuzeichnen. Standardisierte Auszeichnungen hingegen sollen zum einen unabhängig von speziellen Theorien formuliert sein, zum anderen soll jedoch das durch sie zur Verfügung gestellte Inventar von Annotationseinheiten nicht zu umfangreich werden. Wenn Annotationsformate allgemein akzeptiert werden sollen, d. h. wenn sie nicht nur für eine Phänomenklasse oder eine Theorie verwendbar sein sollen, müssen die Formate einen Ausweg aus diesem Konflikt anbieten.

5 Konkrete Annotationsformate

Beispielhaft sollen zwei XML-basierte Annotationsformate vorgestellt werden, die beide zweifelsohne den Status eines Standards besitzen: die Hypertext Markup Language (HTML, ISO:15445, 2000) und die Konventionen der Text Encoding Initiative. Die aktuellen Versionen beider Formate sind XML-konform definiert und erfüllen das Kriterium der Erweiterbarkeit.

5.1 HTML

Es mag überraschend erscheinen, dass in einem Beitrag zur Korpuslinguistik HTML thematisiert wird. In der Tat: Aus linguistischer Sicht bildet HTML kein adäquates Annotationsformat, mehr noch, linguistische Aspekte der Informationsmodellierung fanden zu keinem Zeitpunkt Niederschlag

in den Designkriterien dieses Formates. Konsequenterweise sollte HTML auch nicht im Zusammenhang mit korpuslinguistischen Fragestellungen unmittelbare Anwendung finden. Nichtsdestotrotz kann HTML in der Weise verwendet werden, dass den textuellen Daten linguistische Informationen hinzugefügt werden können. Ein wesentlich wichtigeres Argument bei der Beschäftigung mit HTML besteht allerdings darin, dass die überwiegende Mehrzahl der im WWW zu findenden Informationen in HTML annotiert wurden. Der Korpuslinguistik steht somit eine einzigartige Ressource zur Verfügung (vgl. Rehm, 2002).

HTML ist sehr einfach aufgebaut; HTML-Dokumente haben einen `<head>`, der Information über das Dokument verzeichnet, d. h. die Metainformation, und einen `<body>`, in dem die eigentliche Information enthalten ist, die (normalerweise) im Internet-Browser angezeigt (bzw. bei der Verwendung akustischer Ausgabegeräte vorgelesen) wird. Der im `<body>` enthaltene Text wird in weitere Einheiten (Überschriften, Abschnitte, Listen etc.) strukturiert. Die Annotation von Einheiten, die nicht die Textstruktur betreffen, ist – mit wenigen Ausnahmen – in der strikten Verwendungsweise von HTML nicht zugelassen. Die annotierte Textstruktur bildet eine Wissensquelle, die der Korpuslinguistik von Nutzen sein kann. Beispielsweise finden sich in Überschriften relativ häufig elliptische Konstruktionen.

Informationen zur Textstruktur besitzen die meisten HTML-Dokumente, allerdings gibt es in diesem Format auch weitere Informationen, die für eine linguistische Analyse genutzt werden können, z. B. können die Textteile mittels des Attributes `lang` mit Sprachinformationen versehen werden. Darüber hinaus ist es möglich, mit den allgemeinen Elementen `div` und `span` und dem von allen Elementen nutzbaren Attribut `class` (einfache) linguistische Informationen in HTML-Dokumente zu integrieren. Es bleibt abzuwarten, ob sich eine große Menge von HTML-Dokumenten im Internet findet, die derartige Elemente im ausreichenden Maße verwenden.

Die Weiterentwicklung von HTML, die XML-basierte und modularisierte Version XHTML1.1 (Altheim und McCarron, 2001) erlaubt die Integration verschiedener, separat definierbarer Annotationsformate. Ein derartiges Modul, das Modul *ruby*, wurde bereits vorgestellt, andere frei definierbare XHTML-Module können z. B. der Annotation linguistischer Informationen dienen. Hierdurch wird es möglich, sehr spezielle (z. B. auch selbstdefinierte) Annotationen in HTML-Dokumente zu integrieren.

5.2 Die Text Encoding Initative

Anders als die *Document Type Definition* von HTML war die DTD der *Text Encoding Initiative* (TEI) von Beginn ihrer Entwicklung an modular konzipiert. Die TEI schlägt eine Vielzahl so genannter Tag-Sets vor, die zur Annotation und Repräsentation unterschiedlicher Textsorten verwendet werden können, wobei die linguistische Annotation nicht im Zentrum steht, sondern vielmehr eine von vielen Anwendungen dieses Formates bildet.

Die DTD definiert verschiedene Module, die ihrerseits unterschiedlichen Status besitzen. Die Folge ist, dass einzelne Kern-Elemente in allen TEI-konform annotierten Dokumenten verwendet werden können, die Verwendung anderer Elemente hingegen wird erst durch die Auswahl bestimmter Module möglich.

In den obligatorischen Modulen (den so genannten *core tag-sets*) wird eine allgemeine Dokumentstruktur festgelegt, die der von HTML sehr ähnlich ist, allerdings erlaubt der TEI-Ansatz eine weitaus stärker an den Korpustyp angepasste Annotation, was durch die Auswahl eines sog. *base tag-sets* ermöglicht wird.[5] Ergänzend hierzu können zusätzliche Module eingebunden werden. Diese *additional* bzw. *auxiliary tag-sets* erlauben spezielle Auszeichnungen. So ist es möglich Merkmalsstrukturen zu repräsentieren, die in unifikationsbasierten Grammatikformalismen (vgl. Witt und Müller, 2002) Verwendung finden. Die Annotation morphosyntaktischer Informationen wird ebenfalls durch die Einbindung eines *additional tag-sets* ermöglicht. Die Einbindung dieses DTD-Moduls erlaubt die Auszeichnung von Sätzen (`<s>`), Teilsätzen (`<cl>`)[6], Phrasen (`<phr>`), Wörtern (`<w>`), Morphen (`<m>`) und Zeichen (`<c>`)[7].

Die Möglichkeit zur Erweiterung der TEI-DTD wird durch Mechanismen gewährleistet, mit denen sowohl frei neue Elemente und Attribute definiert werden können, als auch die bereits definierten Elemente veränderbar werden. Diese Möglichkeiten werden z. T. individuell genutzt, es existieren jedoch auch Modifikationen der TEI-DTD, die ihrerseits bereits den Status eines Standards besitzen. Der unter anderem zur Annotation linguistischer

---

[5] Für die Korpuslinguistik sind das *base tag-set* für Prosatexte und das *base tag-set* für die Transkription gesprochener Sprache von besonderer Relevanz.

[6] Der englische Terminus *clause* bezeichnet derartige satzähnliche Einheiten, z. B. Nebensätze.

[7] Zeichen, engl. *characters*, sind normalerweise nicht als linguistische Einheiten zu klassifizieren. Dennoch ist die Verwendung des Elementes `<c>` durchaus auch für sprachwissenschaftlich fundierte Annotationen notwendig, z. B. kann dem Fugenelement 's', das sich in vielen deutschen Komposita findet, nicht der Status eines Morphs bzw. Morphems zugewiesen werden, da es keine Bedeutung trägt.

Korpora geeignete Corpus Encoding Standard (CES) ist ein gutes Beispiel hierfür (vgl. Ule und Hinrichs, in diesem Band).

6 Annotation auf verschiedenen Ebenen

Es wurde bereits erwähnt, dass seit geraumer Zeit die Auszeichnung von Korpora immer häufiger mittels XML-basierter Konventionen bzw. Standards erfolgt. Die breite Verwendung dieses Formates als Annotationsbasis bringt eine Vielzahl von Vorteilen mit sich (Witt, 1998), allerdings ist dieses Format auch nicht frei von Nachteilen. Einem der gravierendsten Probleme von XML wendet sich dieser, den Beitrag abschließende Abschnitt zu.

Texte können auf unterschiedlichen Ebenen strukturiert sein. Renear et al. (1996) diskutieren eine der Grundannahmen der Strukturierung von Textdaten: die „OHCO-These". Diese besagt, dass ein Text eine „ordered hierarchy of content objects" bildet. Diese angenommene Hierarchie rechtfertigt eine grundlegende Einschränkung von XML, die sich selbst in der schwächsten Form der Dokumentannotation – den wohlgeformten XML-Dokumenten – wiederfindet: Durch Elemente ausgezeichnete Umgebungen dürfen sich nicht partiell überlappen. Erfolgt jedoch eine direkte Zuordnung von linguistischen Phänomenen zu der Verwendung von XML-Elementen, zeigt sich, dass derartige Überlappungen ausgesprochen häufig auftreten. Am offensichtlichsten ist der Fall, dass in einem Disput eine Äußerung beginnt, während die andere noch nicht beendet wurde. Aber auch bei der Annotation verschiedener sprachlicher Beschreibungsebenen finden sich häufig derartige Strukturen, z. B. verläuft die Silbenstruktur eines Wortes – wie in Abschnitt 3 (S. 200) gesehen konstituierend für manche Schriftsysteme – oft konträr zu seiner morphologischen Struktur. Nachfolgend werden drei Ansätze dargestellt, die dieses Problem zu lösen versuchen, ohne den durch XML abgesteckten Rahmen zu verlassen.

6.1 Verweise auf eine vorhandene primäre Ebene

Vielfach wurde vorgeschlagen, eine primäre Annotationsebene zur Markierung der wesentlichen strukturellen Einheiten zu verwenden und zusätzliche Ebenen der Annotation mit dieser Ebene zu verknüpfen (vgl. u. a. Barnard et al., 1995; Sperberg-McQueen und Burnard, 1994). Diese Methode wurde auch von dem europäischen Verbundprojekt MATE (vgl. McKelvie et al., 2001) gewählt.

Die Annotation besteht aus verschiedenen Schichten, von denen eine den Status einer primären Ebene erhält. Allen XML-Elementen, die in dieser Annotation verwendet werden, wird ein eindeutiger Identifikator zugewiesen. Diese Elemente bilden dann die (potentiellen) Verweisziele der weiteren Annotationsebenen.

Die Gesamtannotation besteht aus verschiedenen Teilen: der Annotationsbasis und 0 bis n zusätzlichen, mit ihr durch Hypertextverknüpfungen verbundenen Annotationsebenen. Die Ebenen werden so gewählt, dass keine Überlappungen innerhalb einer Ebene auftreten. Überlappungen, die zwischen getrennten Ebenen auftreten, sind – aus der Sicht von XML betrachtet – unproblematisch.

Bei diesem Verfahren ergibt sich ein grundsätzliches Problem: Die potentiellen Beginn- und Endmarkierungen nicht primärer Annotationsebenen müssen bereits in der Basisannotation vorhanden und mit Identifikatoren versehen sein. Davon kann jedoch im Normalfall nicht ausgegangen werden, es sei denn eine Ebene wird zur Basis erwählt, die nicht weiter unterteilt werden kann. Es ist schwer zu entscheiden, welches diese Ebene sein könnte. Neuere Verweistechniken, z. B. XPointer, erlauben zwar eine Lösung dieses Problems, da sie Verweise auf unannotierte Textausschnitte realisieren können, allerdings wird hierdurch die Interpretation der Annotation wesentlich komplexer, sowohl für den Menschen als auch für die Maschine.

Ein weiterer Nachteil dieses Ansatzes besteht darin, dass sich – mit Ausnahme der Basisannotationsebene – die einzelnen Annotationsebenen nicht aus sich heraus erklären. Hinzu kommt, dass die Möglichkeiten von XML als Informationsmodellierungsformalismus nicht ausgenutzt werden. So werden in MATE häufig zur Annotation komplexer Einheiten, die textuelle Daten beinhalten (z. B. Phrasen bzw. *chunks*), in den entsprechenden DTDs leere Elemente definiert.

6.2 Verweise auf Zeitachsen

Eine Teilmenge linguistischer Korpora bilden die Korpora gesprochener Sprache. Diese Gruppe von Daten ist im Normalfall dadurch geprägt, dass es eine primäre Datenebene – die Aufzeichnung der gesprochenen Sprache auf Audio oder Video – gibt. Auf diese Aufzeichnung kann sich mittels eines Zeitstrahls bezogen werden. Die weiteren Ebenen der Datenrepräsentation, Transkriptionen oder Analysen, können als sekundäre Datenebenen angesehen werden, die mit der Zeitachse verbunden sind.

Bird und Liberman (2001) entwickeln einen formalen Rahmen zur Repräsentation sprachlicher Daten, die „Annotation Graphs". Sie bilden ein Modellierungsinstrument, um linguistische Annotationen vorzunehmen.

Ein Annotationsgraph erlaubt die Verknüpfung einer Menge von Knoten, die direkt oder indirekt mit einer *timeline*[8] verbunden sind, mittels benannter Kanten; d. h. Sprachdaten werden so repräsentiert, dass eine primäre Datenebene, die Zeitachse, und eine erweiterbare Menge von Annotationsebenen in einem Formalismus zusammengefasst werden. Alle Ebenen sind mittelbar oder unmittelbar miteinander verknüpft.

Die von Bird und Liberman (2001) hervorgehobenen Ziele des Formalismus lassen sich unter den Kriterien *Einfachheit* sowie *Balance zwischen Generalisierbarkeit und Spezifik* zusammenfassen. Die ausführlich beschriebene Formalisierung der Annotationsgraphen und ihre Anwendbarkeit auf eine immense Anzahl existierender linguistischer Korpora bzw. Annotationsformate stieß in der linguistischen Forschung auf eine äußerst große und mehrheitlich positive Resonanz.

Die Wahl der Verknüpfungsbasis der Annotationsgraphen – die *timeline* – ist jedoch schon vorher als Option für eine möglichst exakte Transkription gesprochener Sprache diskutiert worden. So findet sich in der Dokumentation zur DTD der TEI bereits der Vorschlag zur Verbindung bzw. Synchronisation verschiedener primärer Datenflüsse mittels einer Zeitachse (vgl. Sperberg-McQueen und Burnard, 1994; Johansson, 1995).

Die Verwendung einer Zeit- bzw. einer anderen als primär definierten Datenachse als Referenzbasis bringt jedoch auch einige Nachteile mit sich. Einer der gravierendsten Mängel wurde von Stig Johansson, dem Entwickler des TEI-Moduls zur Transkription gesprochener Sprache, wie folgt formuliert: „It is a problem, however, that the coding becomes rather complex" (Johansson, 1995, S. 156). Es gestaltet sich als äußerst schwierig, die Prozesse der Annotation, Analyse und Darstellung softwaretechnisch unter Rückgriff auf Zeitachsen zu realisieren. Darüber hinaus wird auch in diesem Ansatz von den Möglichkeiten einer XML-basierten Informationsmodellierung kein Nutzen gezogen. So ist es möglich, dass der Beginn und das Ende einer linguistischen Einheit, z. B. eine Phrase oder ein Morph, nur durch Verweise auf die separate Ebene, die *timeline*, definiert wird. Dadurch kann jedoch nicht mehr in einer einfachen Weise angegeben werden, ob das geöff-

[8] Bird und Liberman (2001) weisen darauf hin, dass die Primärdatenachse nicht notwendigerweise ein Zeitstrahl sein muss, welcher die Dauer der Äußerungen festhält. Dies erlaubt es, den Ansatz auch auf linguistische Korpora anzuwenden, die nicht aus einer Transkription gesprochener Sprache resultieren.

nete Morph wieder geschlossen werden muss oder dass Morphe in Phrasen vorkommen können, jedoch Phrasen nicht in Morphen.

Ein weiteres Problem ist ebenfalls in ähnlicher Form bei der Kritik zum MATE-Ansatz beschrieben worden: Eine einzelne Annotationsebene ist nicht separat, d. h. ohne die Basisannotation, verwendbar. Da es auch möglich ist, Querbezüge zwischen den verschiedenen Annotationsebenen aufzubauen (indirekte Verknüpfungen), müssen u. U. alle Ebenen betrachtet werden, auch wenn nur eine einzelne Ebene analysiert werden soll. Neben dem hohen Verarbeitungsaufwand ist es hierdurch de facto unmöglich, die Annotation von nur wenigen Ebenen in einer Art partieller Annotation auszutauschen.

6.3 Separate Annotation

Eine bisher noch äußerst selten betrachtete Möglichkeit zur Dokumentannotation besteht darin, die einzelnen Annotationsebenen vollständig zu trennen. Einzig die TEI-Guidelines erwähnen diese prinzipielle Vorgehensweise, führen sie aber nicht weiter aus (Sperberg-McQueen und Burnard, 1994, S. 755 f.). In einem Teilprojekt der DFG-Forschergruppe „Texttechnologische Informationsmodellierung" kommt dieser Ansatz zur praktischen Anwendung. Das Grundprinzip der separaten Annotation ist denkbar einfach: Jede Annotationsebene wird unabhängig von anderen Ebenen annotiert. Hierbei werden die zu annotierenden Daten entsprechend der Anzahl der Annotationsebenen dupliziert.

Während klassischerweise immer versucht wird, alle zu annotierenden Daten in einer logischen XML-Instanz aufzunehmen, werden hier die Informationen auf mehrere, möglicherweise sehr viele, logische Dateien verteilt. Dies stellt einen fundamentalen Wechsel der Herangehensweise dar.

Die zugrundeliegende Idee besteht darin, dass es nicht notwendig ist, elaborierte Verknüpfungsmechanismen zu entwickeln und die entsprechenden Hyperlinks in die Annotationen einzubauen, sondern die annotierten Primärdaten als Basis der Verknüpfung zu verwenden. Da die annotierten Texte jeweils Kopien der Ausgangsdaten sind, wird für jede der separaten Annotationen derselbe Text verwendet. Die (unterschiedlich) annotierten Texte selbst ermöglichen bzw. bilden die Verknüpfung.

In Sasaki et al. (2002) wird eine Anwendung dieses Verfahrens auf die Annotation von Koreferenz im Japanischen und im Kilivila (einer austronesischen Sprache) vorgestellt. Verschiedene sprachspezifische Mittel zur Realisierung von Koreferenz wie Pronomina, 0-Pronomina oder klassifikatorische

Ausdrücke werden separat annotiert, ebenso Subtypen semantischer Beziehungen wie strikte Koreferenz oder semantische Mengenbeziehungen. Auf diese Weise ist das resultierende Korpus unterschiedlichen Anwendungsgebieten wie Anaphernauflösung oder automatischer Erkennung von Text- oder Diskursstrukturen leicht zugänglich: Die jeweils relevanten Annotationsebenen können in der notwendigen Weise kombiniert beziehungsweise ein- und ausgeblendet werden.

Um den Weg der separaten Annotation zur Repräsentation verschiedener logischer Ebenen beschreiten zu können, müssen jedoch mehrere Beschränkungen in der Datenhaltung beachtet werden. An erster Stelle ist hierbei die konsistente Datenhaltung und die begrenzte Veränderbarkeit der Daten zu nennen. Die zu annotierenden Daten müssen in jeder einzelnen der separaten Auszeichnungsebenen identisch sein. Dies bedeutet insbesondere, dass Veränderungen der Primärdaten in allen Annotationsebenen vorgenommen werden müssen, damit die (impliziten) Verknüpfungen aufrecht erhalten werden können. Vergleichbar hierzu ist das „klassische" Problem, dass Referenzzielpunkte (die mit Attributen vom Typ `ID` gekennzeichneten Knoten) bei ihrer Tilgung besonders überprüft werden müssen, um die Zerstörung potentiell existierender Querverweise zu vermeiden. Eine Reihe weiterer Punkte betreffen die konsistente Interpretation der Daten, etwa bei Leer- bzw. Trennzeichen.

Werden diese Punkte beachtet, könnte die multiple Annotation ein sehr mächtiges und adäquates Mittel zur Repräsentation der Informationen über multiple Annotationsebenen und somit eine geeignete Technik zur Auszeichnung linguistischer Korpora bilden, insbesondere in Hinblick auf den immer mehr an Bedeutung gewinnenden Aspekt der Wiederverwendung (linguistischer) Ressourcen.

Literaturverzeichnis

Altheim, Murray und McCarron, Shane (2001): "XHTML 1.1 – Module-based XHTML". `http://www.w3.org/TR/xhtml11/`.

Aston, Guy und Burnard, Lou (1998): *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Barnard, David T.; Burnard, Lou; Gaspart, Jean-Pierre; Price, Lynne A.; Sperberg-McQueen, C. M. und Varile, Giovanni Battista (1995): "Hierarchical Encoding of Text: Technical Problems and SGML Solutions".

In: *Text Encoding Initiative: Background and Context*, herausgegeben von Ide, Nancy und Véronis, Jean, Dordrecht: Kluwer, S. 211–231.

Bird, Steven und Liberman, Mark (2001): "A formal framework for linguistic annotation". *Speech communication* 33 (1, 2): S. 23–60.

Burger, Susanne; Weilhammer, Karl; Schiel, Florian und Tillmann, Hans G. (2000): "Verbmobil Data Collection and Annotation". In: *Verbmobil: Foundation of Speech to Speech Translation*, herausgegeben von Wahlster, Wolfgang, Berlin: Springer.

Erjavec, Tomaz; Lawson, Ann und Romary, Laurent (1998): "East meets West: Producing Multilingual Resources in a European Context". In: *Proceedings of the First International Conference on Language Resources and Evaluation (LREC'98)*. Granada.

Gippert, Jost (1999a): "Language-specific encoding in multilingual corpora: Requirements and solutions". In: *Multilinguale Corpora – Codierung, Strukturierung, Analyse*, herausgegeben von Gippert, Jost, Prag: Enigma, S. 372–384.

Gippert, Jost (Herausgeber) (1999b): *Multilinguale Corpora – Codierung, Strukturierung, Analyse*. Prag: Enigma.

ISO:10646 (2000): *ISO/IEC 10646: Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Version 3.0. Part 1: Architecture and Basic Multilingual Plane, Part 2: Supplementary Planes (The Unicode Standard)*. International Organization for Standardization, Genf.

ISO:15445 (2000): *ISO/IEC 15445: Information technology – Document description and processing languages – HyperText Markup Language (HTML)*. International Organization for Standardization, Genf.

ISO:639 (1998): *ISO/IEC 639: Codes for the representation of names of languages. Part 1 (1988), Part 2 (1998)*. International Organization for Standardization, Genf.

ISO:8859-1 (1998): *ISO/IEC 8859-1: Information technology – 8-Bit single-byte coded graphic character sets. Part 1: Latin alphabet No. 1*. International Organization for Standardization, Genf.

ISO:8859-2 (1999): *ISO/IEC 8859-2: Information technology – 8-Bit single-byte coded graphic character sets. Part 2: Latin alphabet No. 2*. International Organization for Standardization, Genf.

ISO:9541-1 (1991): *ISO/IEC 9541-1: Information technology – Font information interchange. Part 1: Architecture*. International Organization for Standardization, Genf.

Johansson, Stig (1995): "The Encoding of Spoken Text". In: *Text Encoding Initiative: Background and Context*, herausgegeben von Ide, Nancy und

Véronis, Jean, Dordrecht: Kluwer, S. 149–158.

Karlsson, Fred; Voutilainen, Atro; Heikkilä, Juha und Anttila, Arto (Herausgeber) (1995): *Constraint Grammar*. Berlin: de Gruyter.

McEnery, Tony und Wilson, Andrew (1996): *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McKelvie, David; Isard, Amy; Mengel, Andreas; Møller, Morten B.; Grosse, Michael und Klein, Marion (2001): "The MATE Workbench – An Annotation Tool for XML Coded Speech Corpora". *Speech Communication* 33 (1–2): S. 97–112.

McWhinney, Brian (2000): *The childes project: tools for analyzing talk*. Hillsdale (New Jersey): Erlbaum.

Milde, Jan-Torsten und Gut, Ulrike (2002): "The TASX-environment: an XML-based toolset for time aligned speech corpora". In: *Proceedings of the third international conference on language resources and evaluation (LREC 2002)*. Gran Canaria.

Musgrave, Simon (2001): "A brief description of the Spinoza typological database". In: *Proceedings of the IRCS Workshop on Linguistic Databases*, herausgegeben von Bird, Steven; Buneman, Peter und Liberman, Mark. Philadelphia: University of Pennsylvania.

Rehm, Georg (2002): "Schriftliche Mündlichkeit in der Sprache des World Wide Web". In: *Kommunikationsform E-Mail*, herausgegeben von Dürscheid, Christa und Ziegler, Arne, Tübingen: Stauffenburg, S. 263–308.

Renear, Allen; Mylonas, Elli und Durand, David (1996): "Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies". In: *Selected papers from the ALLC, ACH Conference: Christ Church, Oxford, April 1992*. Oxford: Clarendon Press.

Sasaki, Felix; Wegener, Claudia; Witt, Andreas; Metzing, Dieter und Pönninghaus, Jens (2002): "Co-reference annotation and resources: a multilingual corpus of typologically diverse languages". In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*. Las Palmas.

Sperberg-McQueen, C. M. und Burnard, Lou (Herausgeber) (1994): *Guidelines for Electronic Text Encoding and Inter-change (TEI P3)*. Chicago, Oxford: Text Encoding Initiative.

Suignard, Michel; Ishikawa, Masayasu; Dürst, Martin; Texin, Tex und Sawicki, Marcin (2001): "Ruby annotation". W3C Recommendation. `http://www.w3.org/TR/2001/REC-ruby-20010531/`.

Whistler, Ken und Davis, Mark (2000): "Character Encoding Model. Unicode". Technischer Bericht, Unicode, Inc. `http://www.unicode.org/`

unicode/reports/tr17/.

Witt, Andreas (1998): "TEI-based XML-Applications: Transcriptions". In: *ALLC/ACH98, Joint Conference of the ALLC and ACH*. Debrecen.

Witt, Andreas und Müller, Stefan (2002): "Grundlagen für den Computereinsatz in der Linguistik: Attribute, Werte, Unifikation". In: *Arbeitsbuch Linguistik*, herausgegeben von Müller, Horst, UTB.